Cont a1

of a molecule and perform what is known as principal component analysis (PCA). They do this to find representatives of all possible conformations that are most distinct. Although this procedure is really akin to finding the dimensionality of the space in which these conformers exist, Mestres et al. do not use PCA for this purpose, but merely to cluster the conformers. They do not apply PCA to sets of different molecules, only to conformers of the same molecule, and they do not use any other "metric" property of their similarity measure. In fact they seem unaware of such.

Replace the paragraph beginning at page 9, line 20, and carrying over to page 10, line 12 with:

a2

A metric distance may also be used in a technique called "embedding". The number of links between the elements of a set of N elements can be shown to be N*(N-1)/2 and each link can be shown to be a metric distance. While a set of N elements has N*(N-1)/2 distances, the set can always be represented by an ordered set of (N-1) numbers, i.e. I can "embed" from a set of distances to a set of N positions in (N-1) dimensional space. This is identical to Principal Component Analysis mentioned previously, except that with PCA one finds the most "important" dimensions, i.e. the "principal" directions, which carry most of the variation in position. Typically with PCA one truncates the dimensionality at 2 or 3 for graphical display purposes. In general, the number of dimensions which reproduces the set of N*(N-1)/2 distances within an acceptable tolerance may be much smaller than (N-1), yet still be greater than 2 or 3. Hence one talks of "embedding into a hyper-dimensional subspace", where hyper-dimensional means more than 3 dimensions, and subspace means less than (N-1). Techniques for such an embedding are standard linear algebra. When applied to molecular fields,

Cont'd A2

the result of embedding is a shape-space of $M \leq N-1$ dimensions.

Replace the paragraph beginning at page 21, line 26, and carrying over to page 22, line 8 with:

A3

Various techniques exist to attempt to find the best overlap of two fields, typically involving repeated searches from different starting orientations of the two molecules. This is necessary because no direct solution for the minimal distance orientation is available, and most methods tend to get caught in nearby local minima, missing the global minimum. One such technique is a Gaussian technique described in J.A. Grant et al., "A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape," J. Computational Chemistry, Vol. 17, No. 14, pp. 1653-66 (1996). Using this technique, I overlaid the two molecules shown in Fig. 2A to produce the result shown in Fig. 2B.

Replace the paragraph beginning at page 22, line 21 with:

A4

In the following, I may refer interchangeably to maximal overlap (or overlay), minimal field difference, and minimal distance, as they all refer to and measure the same optimal orientation of two molecules with respect to each other.

Replace the paragraph beginning at page 23, line 11 with:

A5

In all of my methods for using the field metric, the steric field for each molecule is constructed either from a sum of Gaussians centered at each atom, or as one minus the product of one minus each such Gaussian. These are referred to as the "sum form" and "product form" respectively. The product form has the advantage that it removes excess internal overlap and hence is smoother inside. The sum form has the advantage that it is numerically simpler. Each Gaussian is such that its volume is the same as that of the atom it contd a5

represents, and the volume, as of any field, is calculated from the integral of the function over all space.

Replace the paragraph beginning at page 26, line 11 and carrying over to page 27, line 11:

a6

For example, if I have 1000 molecules in my database I might organize this information thus: select 10 "key" molecules which are quite different in shape. For each of these 10 key molecules I then find the distance from each of these molecules to every other molecule in the database, and make 10 lists where each list has a different key molecule at the top and the rest of the 999 molecules are listed in order of shortest distance from it. To find the closest match between a test molecule and the 1000 molecules of the database I begin by determining the metric distances between the test molecule and each of the key molecules. Suppose the shortest distance is to key molecule 6 and that distance is X. I now begin to calculate the distances to the rest of the molecules, but in the order specified by that key molecule's list. Since the list has molecules close to key molecule 6 first, it is likely these are also close to my test molecule. Furthermore, by the triangle inequality, since molecules which are a distance greater than 2X from key molecule 6 must be greater than X from my test molecule, I only have to go down the list until this condition is satisfied, i.e. I may not have to test all 1000 molecules. Furthermore, if I find a molecule closer than key molecule 6 early in the list, say distance X-d, then I only have to go down the list until the distance from the key molecule is greater than 2(X-d), i.e. I can refine the cutoff distance as I progress down the list. Thus I can search the database, by shape, in a time sublinear with the number of molecules in the database. These methods are not possible without evaluating a shape space description of the set of molecules that comprise the database.

Replace the paragraph beginning at page 29, line 6 with:

a) Choose the number of EGFs that I want to represent the field.

Replace the paragraph beginning at page 29, line 8 with:

b) Choose random positions for the center of each EGF and make each spherical, i.e. a=b=c=1.

Replace the heading beginning at page 35, line 1 with:

1: **Finding the maximal overlap (minimal field difference) between two fields A and B**

Replace the sub-heading at page 35, line 4 with:

A) Exhaustive Search:

Replace the paragraph beginning at page 42, line 6 and carrying over to page 43, line 3 with:

Once I have a shape space for N molecules, of dimension M, the next step is to calculate the position within this shape space for a molecule not used in the construction of that shape space. This position is found by analogy with triangulation in three dimensions, i.e. if one has a set of distances from an object to four reference objects the exact position can be ascertained. In two dimensions one needs three distances. In M dimensional shape space one needs M+1 distances. (In each of these cases, the M+1 distances must be from points which cannot as a set be described at a dimensionality less than M, e.g. for the case of three dimensions, the four reference points cannot all lie in a 2 dimensional plane). The actual procedure for going from distances to a position is simply that a linear equation for the coordinates can be generated from each distance, such that the solution of the set of such produces the position. This set of linear equations can be solved by a.

any standard method, for instance, Gauss-Jordan elimination (see, for example Stoer and Bulirsch, "Introduction to Numerical Analysis", 2nd Ed., Springer-Verlag, chapter 4). An important note here is that this procedure can fail, i.e. it will produce a position which will underestimate the M+1 distances by a constant amount. This is an indication that the structure under study actually lies in a higher dimensional space than the shape space previously constructed. As such, that shape space needs to be extended.

A9

Replace the paragraph beginning at page 46, line 1 with:

A10

(i) Choose a structure at random from the N possible structures.

Replace the paragraph beginning at page 47, line 19 with:

A11

(ii) From the set of N structures, select K key structures that are quite different from each other (i.e. are remote from each other in shape space). For instance, the structures may simply be different from each other in total volume, or be chosen by more computationally intensive methods, e.g. as representatives of clusters of molecular shapes found by standard clustering techniques (e.g. Jarvis-Patrick, etc). These more sophisticated methods may be greatly speeded if the shape space has been determined.

Replace the paragraph beginning at page 49, line 21 and carrying over to page 50, line 3 with:

A12

Thus I can search the database, by minimum field difference, in a time sublinear with the number of molecules in the database. This is because, by the triangle inequality, I know the cutoff distance for evaluating structures in the list is at most equal to 2X (when BEST = X) and is potentially further refined as I progress down the list and find better a12 (smaller) values for BEST. As noted above, the list creation process can be speeded if the shape space of the structures has already been determined. Whether the time saved will be justified by the time spent constructing the shape space depends on the number of key structures K and the number of structures in the database.

Replace the paragraph beginning at page 50, line 20 with:

a13 (ii) Choose a structure at random from this set and record its name in the zero level node of a tree structure which is such that each "node", or "slot", has two child nodes, called "left" and "right", at what I refer to as a level one greater than this node.

Replace the paragraph beginning at page 52, line 10 with:

a14 In (1) above, rather than choosing structures at random for insertion into the tree, they could instead be sorted into a list, for example in order of increasing volume, and then taken sequentially from the list for insertion into the tree. This allows additional criteria to be used to terminate a search of the branches of the tree.

Replace the paragraph beginning at page 58, line 19 with:

a15 (vi) If the number of EGF's used in (ii) is greater than one check to see if this fragment adjusted EFF is greater than BEST. If so then quit the procedure, otherwise increment the number of EGF's to be used in (ii) by one and return to (ii).

a16 **Replace the paragraph beginning at page 60, line 7 with:**

(iv) For each of the four alignments, make the atom to atom assignments for the atoms which belong to the pair of EGF's being aligned together based upon "closest" or "closest of similar type".

Replace the paragraph beginning at page 60, line 12 with:

a16

(v) Rather than have an infinite number of possible alignments I now have just four to choose from, and given any kind of measure for the assignment (e.g. minimize the sum of the distances of each atom pair) this is straightforward.

Replace the paragraph beginning at page 67, line 6 with:

a17

(ix) Otherwise actually find the best metric field difference between the new molecule and the current database structure. If this value is less than BEST, set BEST equal to this value, set the value of BESTSTRUCTURE to indicate this structure. Go to (v) unless this is the last structure in the database.

Replace the paragraph beginning at page 70, line 11 with:

a18

(i) Define a fitting function f between any two EGF’s such that if both were spherical this function would be a minimum when the inter-EGF distance is the same as the sum of the radii of each EGF (defining the radii of the EGF as that of a sphere of equivalent volume). Such a function for two EGF's, EGF1 and EGF2, is:

f = a*V - b*(Q (EGF1, V) - Q (EGF2, V))
where V = Q (EGF1, EGF2) where Q is defined in equation (6) above.

Replace the paragraph beginning at page 71, line 29 and carrying over to page 72, line 2 with:

a19

This procedure produces a series of single EGF descriptions of the active site. These EGF’s may be painted, based upon properties of the nearest protein atoms, or of any field quantity generated by such atoms, e.g. electrostatic potential.